DOCUMENT RESUME

ED 189 112 TM 800 275

AUTHOR Reynolds, Cecil R.
TITLE Differential Predictive Validity of a Preschool Battery Across Race and Sex.
PUB DATE Apr 80
NOTE 28p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Grade 1; Kindergarten Children; *Predictive Validity; *Preschool Tests; Primary Education; Racial Bias; Sex Bias; *Test Bias; Testing Problems
IDENTIFIERS Lee Clark Reading Readiness Test; McCarthy Scales of Childrens Abilities; Metropolitan Achievement Tests; Metropolitan Readiness Tests; Preschool Inventory (Caldwell); Test of Basic Experience

ABSTRACT

Determination of the fairness of preschool tests for use with children of varying cultural backgrounds is the major objective of this study. The predictive validity of a battery of preschool tests, chosen to represent the core areas of preschool assessment, across race and sex, was evaluated. Validity of the battery was examined over a 12-month period, involving 322 preschoolers (white and black, female and male). A regression equation was determined using all preschool measures to predict achievement scores. Predictions were made for each individual test, and residual terms were calculated. No significant differences occurred in mean residuals between any pair of groups, indicating an absence of bias in prediction across race and sex with the large battery. Sex bias in prediction was seen when subsets of the larger battery were examined. It is suggested that test developers become aware of the issue of bias and demonstrate differential construct and predictive validity as a part of the development of a test. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189112

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Differential Predictive Validity of A Preschool Battery Across Race and Sex

Cecil R. Reynolds
Acting Director
Buros Institute of Mental Measurements
135 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, Nebraska 68588

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. Reynolds

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A paper presented to the annual meeting of the American Educational Research Association, Boston: April, 1980.

TM 800275

Abstract

The study evaluated the predictive validity of a battery of preschool tests, chosen to represent the core areas of preschool assessment, across race and sex. Predictive validity of the battery was examined over a 12 month period utilizing 322 preschoolers (90 white females, 86 white males, 73 black females, 73 black males). The preschool battery was administered at the end of kindergarten and the Metropolitan Achievement Test at the end of first grade. A regression equation was determined using all preschool measures to predict achievement scores. Predictions were made for each individual test, and residual terms were calculated. Residuals were submitted to a race by sex ANOVA for each subtest of the Metropolitan Achievement Test. No significant differences occurred in mean residuals between any pair of groups, indicating an absence of bias in prediction across race and sex with the large battery. When subsets of the larger battery were examined, sex bias in prediction was seen, indicating significant underprediction of female performance in some achievement areas. However, the magnitude of the effect was small.

The past decade has witnessed a substantial surge of interest in preschool assessment, and in the differential validity of psychological tests when used with individuals of varying cultural backgrounds. The use of tests normed primarily on white children is of special concern to educators due to the specifications of P. L. 94-142 (Education for All Handicapped Children Act of 1975) and the now famous Larry P. case in San Francisco[1]. Harrington (1975) has gone so far as to state that it is not possible for tests developed and normed on a white majority to be other than biased against minorities, or to predict accurately when used with minorities.

In reporting on the use of educational and psychological tests with disadvantaged students, a committee appointed by the American Psychological Association, Board of Scientific Affairs (Cleary, Humphreys, Kendrick, & Wesman, 1975) offered a definition of test bias. While including content and construct validity as important variables in the issue of bias, the emphasis was clearly on predictive validity: "A test is considered fair for a particular use if the inference drawn from the test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group" (p. 25). The definition of bias offered by the APA committee is a restatement of previous definitions by Cardall and Coffman (1964), Cleary (1968), Pott-

hoff (1966), and others, and has been widely accepted (though certainly not without criticism, e.g., Bernal, 1975; Thorndike, 1971).

Oakland and Matuszek (1977) examined class placement procedures under several proposed models of bias and demonstrated that application of the Cleary model results in the smallest number of children being misplaced, though under certain legislative conditions they favored the Thorndike selection model. For the present study, the Cleary et al. definition is stated in somewhat different (though reasonably equivalent) terms: A battery of tests is considered fair if it significantly increases one's ability to predict the criterion variable (i.e., $\underline{R} > 0$) and if errors in prediction are independent of group membership (i.e., $\underline{\bar{X}ej}_1 = \underline{\bar{X}ej}_2 = \ldots \underline{\bar{X}ej}_k$).

Much research is available on the validity of the SAT for blacks and whites. A few studies dealing with the predictive validity of IQ tests with elementary and secondary school children have recently appeared (e.g., Reynolds & Hartlage, 1979). However, only a single study has been located dealing with the differential validity of preschool tests. Mitchell (1967) studied the validity of two broad based readiness tests used to predict first grade achievement for blacks and whites, and found similar validity coefficients for the two races. However, Mitchell's study was limited to comparing the magnitude of the independent-dependent variable relationship and did not look at

error terms when a common equation is applied to all groups.

Since most major preschool assessment devices (e.g., Wechsler Preschool and Primary Scale of Intelligence, McCarthy Scales of Children's Abilities) require highly trained personnel and more than one hour of testing time per child, they are typically not employed in large scale screening programs. Consequently, a variety of short screening tests and several group readiness tests have been developed and normed for the preschooler. Available instruments include tests of basic concepts, criterion referenced "Achievement" tests, tests of visual-motor integration and nonverbal concept formation, and general readiness screening batteries (Reynolds, 1979). The present investigation includes tests representative of each of these skill areas. These tests are frequently used in placement decisions (e.g., regular first grade versus a developmental classroom), deciding upon referrals for more comprehensive psychological evaluation, and in decisions on early admissions to kindergarten or first grade. The study is designed to determine the fairness of preschool tests for use with children of varying cultural backgrounds. More specifically, the following question is asked: When preschool tests representative of the core areas of preschool assessment are grouped into a battery used to predict first grade achievement, are the errors inherent in statistical prediction (residual terms) randomly distributed by race and sex when a common multiple regression equation is applied to all scores?

METHOD

Subjects

The initial sample for pretesting consisted of 393 kindergarten children from a school district located in a small metropolitan area in the southeastern United States. Complete data were obtained on 322 of the 393 children, representing an attrition rate of 18% over the one year period. The final sample contained 90 white females (WF), 86 white males (WM), 73 black females (BF), and 73 black males (BM). The total group had a mean age of 82.57 months at the time of post-testing (end of first grade), with no race or sex differences by age. The sample was primarily middle to lower-middle class in SES ratings, with the white SES rating significantly higher than the blacks' (t = 29.5, df = 320, $p \leq .001$). The children are described in greater detail in Table 1 and elsewhere (Reynolds, 1978a).

Insert Table 1 about here

Test Instruments

Pretesting was conducted in May of the kindergarten year on each of the following tests, with the exception of the Metropolitan Readiness Test which was administered during the second week of first grade. Pretests included:

1) The McCarthy Drawing Tests: Draw-A-Design (DAD) and Draw-A-Child (DAC). DAD and DAC are subtests 12 and 13 of the

McCarthy Scales of Children's Abilities (McCarthy, 1972). The two tests measure primarily visual-motor integration and non-verbal concept formation. Although designed to be individually administered, Reynolds (1978b) found that the two tests can be validly administered in a group setting with the obtained scores showing significant correlation (typically around .50) with measures of achievement. Reynolds also obtained test-retest reliabilities from group to individual administration for DAD and DAC of .86 and .82, respectively. In the present study, the two tests were administered in groups of 8 to 10 by a kindergarten teacher and aide trained in the administration by the investigator. Scoring was done by a doctoral level school psychologist with no knowledge of the child's race or age.

2) Lee-Clark Reading Readiness Test (Lee & Clark, 1962): The Lee-Clark is a group administered reading readiness test which requires the matching and differentiation of letters, basic concept recognition, and letter and word identification. The Manual reports split-half reliabilities, corrected by Spearman-Brown, of .96 for two separate samples of kindergarten pupils (N's = 94 and 80). Validity coefficients hovering about .50 with several first grade reading tests are also reported in the Manual.

3) Tests of Basic Experiences (TOBE; Moss, 1970): The TOBE are a series of five standardized group tests for young children. For the present study, two of the tests are administered: Mathematics and Language. The TOBE Mathematics subtest, according

to the author, measures the child's mastery of basic math concepts and the terms associated with them (e.g., biggest, oldest, most, etc.). The TOBE Language test primarily deals with vocabulary, sound-symbol relationships, listening skills, and letter recognition. The TOBE Manual reports KR_{20} reliability estimates of .80 for Mathematics and .84 for Language with a sample of 700 kindergarten children. Only evidence of content validity is given in the TOBE Manual, which is based on teacher judgments as to the correct classification of each item (i.e., Math, Language, Science, etc.). A high percentage of agreement was obtained between teachers' classifications of the items and the items' actual classification.

4) Preschool Inventory - Revised Edition (Educational Testing Service, 1970): The Preschool-Inventory is divided into four subtests measuring Personal-Social Responsiveness, Associative Vocabulary, Concept Activation-Numerical, and Concept Activation-Sensory. The Inventory is individually administered and consists of a series of general information questions and simple activities such as design copying. The Inventory was carefully normed according to U.S. Census data. KR_{20} reliability estimates ranging from .86 to .92 are reported for the standardization sample at various age ranges with a total sample KR_{20} of .91. Correlations with the Stanford-Binet Intelligence Scale are reported in the Manual and range from .39 to .65 with a total sample correlation of .44. The sample consists of 1,476 children

divided into five age groups between 3-0 and 6-11 for norming and other statistical treatments. For the present study, the inventory was administered individually to each student and scored by a State of Georgia Department of Education certified school psychologist.

5) Metropolitan Readiness Tests (Hildreth, Griffiths, & McGauvran, 1969). The MRT were devised as a group test to measure the various skills in young children which contribute to readiness for first grade, and according to the authors, provide a "quick, convenient, and dependable basis for the early classification of pupils . . ." Six basic subtests and one alternate, an adaptation of the Goodenough-Harris Drawing Test, are provided measuring vocabulary, visual-motor integration, verbal comprehension, letter identification, and a variety of other subskills (e.g., visual perception and discrimination). Split-half reliabilities are reported in the MRT Manual and range from .90 to .95 for the total test with seven different samples. Subtests' reliabilities are considerably lower ranging from .33 to .89 across the seven samples. Alternate form reliabilities range from .89 to .93 for total test and .50 to .86 for individual subtests. Numerous validity studies are reported in the Manual and show correlations with later achievement measures consistently hovering about .50. Due to the results of several factor analytic studies (Reynolds, 1979; in press), only the MRT total test score was used in the present study.

The criterion measure chosen was the Metropolitan Achievement Tests, Primary I Battery (MAT). The MAT was chosen for several reasons. Within the school system in which data were collected, the MAT had been chosen by the first grade teachers from a selection of standardized first grade achievement tests for five consecutive years as the test which most closely measured what they taught. While this consideration is of primary importance, Anastasi (1976) also singles out the MAT as an exemplary model of an achievement test designed to eliminate or reduce racial bias in item content. The MAT is a group administered test yielding subscores for Word Knowledge, Word Discrimination, Reading, and Arithmetic. Split-half reliabilities for the subtests range from .81 to .95 with a median reliability of .91.

Procedure

During the last two weeks of May, 393 of approximately 400 kindergarten children were administered the two McCarthy Drawing Tests, the two TOBE subtests, the Lee-Clark Reading Readiness Test, and the Preschool Inventory. The following fall during the first two weeks of school, all first grade children were administered the MRT. In May of the first grade year, all students were administered the MAT. Any student with missing data for any test was eliminated from all analyses, leaving 322 children.

Data Analysis

Since the age range of the sample was nearly two years, all

raw scores were converted to standard scores based on data available in their respective manuals. Due to a lack of information concerning the raw score distributions of the MRT and the MAT, scores from these two tests were converted to percentile ranks. A multiple regression equation was computed employing all independent variables in the prediction of MAT scores for the total sample. An equation was calculated for each MAT subtest. Including all independent variables in the multiple regression formula, instead of using a stepwise procedure when examining for bias, increases reliability (Linn & Werts, 1971) and lessens the probability of finding artifactual differences. Using the equations for N=322, predicted MAT scores were obtained for each subtest. Residuals were then calculated. To determine whether residuals were randomly assigned by race and sex, a 2 x 2 ANOVA (race by sex) was calculated on the residual terms, producing four separate analyses (one for each MAT subtest), the null hypothesis in each case being: H_0: $\mu_{e_{BM}} = \mu_{e_{WM}} = \mu_{e_{BF}} = \mu_{e_{WF}}$. Use of the ANOVA approach to this question allows for examination of the race by sex interaction in the prediction of first grade achievement from preschool screening test scores, and is a straight-forward test of the Cleary et al. (1975) condition that "there is no constant error in the inference as a function of membership in a particular group." A signficant F - ratio in this case indicates nonrandom error and thus bias in prediction. Since four separate ANOVAs were calculated, a significance level

of $p \leq .01$ was adopted.

As an internal replication of the study, two of the pretests were chosen (TOBE-Language and the Lee-Clark) to be examined in combination without including the other pretests. Regression equations for each dependent variable with the TOBE-Language and Lee-Clark as predictors were calculated for N = 322 and the analysis of residuals repeated as described above. Since most school districts are likely to apply only one or two screening tests (instead of the large battery of the present study), this analysis allows examination of the more practical implications of the present study. One further analysis was conducted to also aid the generalizability of results across tests and to practical screening settings.

The MRT is certainly one of the most widely employed screening tests in the public schools (Reynolds, 1979; in press), frequently being the only screening measure employed. In the current study, the MRT showed the largest zero-order correlation with the dependent measures. For these reasons, a regression analysis of predictive bias by the MRT was undertaken via the Potthoff (1966) procedure. With a single independent variable, the Potthoff procedure is a computationally simple comparison of regression lines that provides a single F-ratio simultaneously testing slope and intercept values. If a significant F results, separate tests of slopes and intercepts may then be conducted to determine if the bias in prediction is constant (intercepts differ) or changes with distance of the score from

the mean (slopes differ). The results of the Potthoff analysis are interpreted consistent with the analysis of residuals since if regression lines are simultaneous across groups, residuals must also be equal.

RESULTS AND DISCUSSION

The mean performance for each subgroup within the total sample of 322 is discussed in detail in Table 1. As a general rule, performance on the pretests and posttests rank ordered WF>WM>BF>BM, with the notable exception of the McCarthy Draw-A-Child subtest which ordered WF>BF>WM>BM.

Summaries of the results of the four ANOVAs for the total battery are presented in Table 2. No significant differences in

---

Insert Table 2 about here

---

the average errors of prediction for any of the dependent variables occurred by race or sex and no race by sex interactions proved significant at the .01 level of probability. Thus, the mean error of prediction for all race/sex groupings was essentially equivalent and approached zero in all cases. Had the $p \leq .05$ level been adopted, a single significant finding would have resulted, showing slight underprediction of performance for females on the MAT - Word Knowledge subtest.

In the multiple regression equation for each of the dependent measures, the MRT maintained the largest Beta weight in all cases except one, giving it the greatest influence in the prediction

equation. As seen in Table 3, Rs were in the high .60s and low .70s. The MRT also had the largest zero-order correlation with each criterion variable when compared to the other pretest scores. While this may be a function of the large number of abilities tapped by the MRT, it must be recalled that the MRT was administered three months closer to the criterion measures. The actual regression equations are reported in Table 3.

Insert Table 3 about here

When the MRT was examined alone as a predictor of first grade academic attainment, no bias in prediction occurred for any group. None of the F tests for differences in regression lines between groups even approached statistical significance ($p \leq .01$) for any dependent variable. Fs ranged only from 0.15 to 3.16.

Some differences did occur when only the TOBE-Language and Lee-Clark were used as predictors however. Summaries of the results of the four ANOVAs for the two-test regression equations are presented in Table 4. No main effects occurred for race and

---

Insert Table 4 about here

---

there were no significant race by sex interactions for any of the dependent measures. With regard to sex, main effects occurred for the MAT-Word Discrimination and Word Knowledge

subtests but not for Reading or Arithmetic. Examination of the mean residuals for each group for Word Discrimination and Word Knowledge showed consistent underprediction of performance by females on each measure. The magnitude of the effect was small however, being on the order of .16 standard deviations on Word Discrimination and .13 on Word Knowledge. This translates into approximately 2-3 percentiles. This finding is consistent with a trend noted in the analysis of the larger battery. These two screening tests would tend to over identify females as potential candidates for remedial or special education classes. However, it is interesting to note that special education classes as a rule are more sexist than racist in makeup with boys outnumbering girls 4:1. While the finding of sex bias in prediction is interesting and indicates a real need for follow-up study with a variety of other tests, the small magnitude of the effect, specificity of appearance (occurring only in two rather limited areas of achievement), and the relatively small percentage of females in special education programs indicate few problems of practical significance in the use of these tests in preschool screening.

The present study does not indicate the presence of bias in the prediction of first grade achievement across race and sex when a large battery of tests is employed. Racial bias in prediction also did not occur for the single test nor the two-test battery. Sex effects with the two-test battery suggest a need for testing a diverse sample of skills to avoid bias in predic-

tion. The MRT tests a large number of abilities as does the larger battery; the two-test battery employed is more specific to reading and language related skills. While testing a wide range of abilities may help insure against bias in prediction, the small magnitude of the effect would not support the use of more than several tests under the scrutiny of any reasonable cost-benefit analysis. Use of the MRT as an integral part of a screening program seems to be one of the most reasonable alternatives. However, much replication and further research will be needed before the effects of race and sex on our ability to predict early academic status if fully understood. The researcher and practitioner should be aware of the potential for bias existent in working with preschool populations. Particular attention should be awarded the sex variable. It has been long ignored in studies of differential validity of cognitive batteries, although it has received the attention of personality researchers, at least with regard to construct validity (e.g., Katzenmeyer & Stenner, 1977; Ozehosky & Clark, 1971).

Test developers also need to become aware of the issue of bias to the point of demonstrating differential construct and predictive validity as a part of the development of a test. Test authors and publishers need to demonstrate factorial invariance across all groups for whom the instrument is designed, in order to make the instrument more readily interpretable. Comparisons of predictive validity across race and sex during the test develop-

ment phase are also desirable. At this stage, tests can be altered through item analysis techniques to eliminate racial and/or sexual bias in the prediction of performance. Studies similar to the present investigation are needed with existing measurement instruments to determine whether alterations in interpretation are needed when applied to various populations. A variety of criteria also need to be employed, including other group achievement measures, individual tests of academic attainment, and teacher-made tests. This is true for the affective as well as the cognitive domain (Reynolds, 1978a).

Table 1

Means and SD's for each race/sex grouping on all pretests, posttests, age, and SES.

Variable Name	Group Identification: White Females		White Males		Black Females		Black Males	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
TOBE-Language	111.21	11.23	106.67	12.72	88.97	14.48	90.58	12.54
TOBE-Math	106.84	10.48	105.10	11.22	88.19	12.61	88.81	16.99
Preschool Inventory	119.43	6.88	117.43	11.28	105.48	11.65	104.19	12.14
Lee-Clark	114.21	8.83	106.28	18.40	93.22	21.40	92.48	22.30
Draw-A-Design	93.61	15.31	91.22	14.73	88.08	13.84	78.84	13.08
Draw-A-Child	103.89	18.28	96.40	15.84	98.90	17.82	85.41	16.26
MRT-Total Test	67.59	14.16	62.59	17.82	50.23	14.89	44.95	14.20
MAT-Word Knowledge	69.61	25.64	59.26	32.10	51.44	26.11	38.30	24.94
MAT-Word Discrimination	69.63	24.30	60.98	27.38	53.85	26.18	44.49	27.40
MAT-Reading	66.11	28.69	56.71	31.32	44.37	32.47	36.77	25.68
MAT-Arithmetic	71.91	19.99	67.12	20.45	53.41	19.99	49.14	19.51
SES	2.42	1.14	2.52	1.15	3.58	1.04	3.73	0.98
Age (in months)	82.56	3.60	82.59	3.87	82.58	3.57	82.53	3.98

Table 2

Analysis of Variance Summary Tables for Analysis of Residuals Based on All Independent Variables as Predictors

I MAT - Word Discrimination

Source	df	MS	F*
Race	1	721.14	1.58
Sex	1	462.62	1.02
Race x Sex	1	138.13	0.30
Error	318	454.79	

II. MAT - Word Knowledge

Source	df	MS	F*
Race	1	326.86	0.61
Sex	1	2629.52	5.48
Race x Sex	1	350.92	0.73
Error	318	480.08	

III. MAT - Reading

Source	df	MS	F*
Race	1	749.87	1.49
Sex	1	137.71	0.27
Race x Sex	1	5.17	0.01
Error	318	504.46	

IV. MAT - Arithmetic

Source	df	MS	F*
Race	1	77.30	0.31
Sex	1	167.99	0.68
Race x Sex	1	287.80	1.17
Error	318	246.36	

* none significant at $p \leq .01$

Table 3

Multiple Regression Equations for the Prediction of MAT Scores Employing All Independent Variables

Dependent Variable	R	Prediction Equation
MAT - Word Knowledge	.67	.72 (MRT)* = .33 (TOBE-Language)* + .12 (DAC)* + .09 (Lee-Clark)* + .09 (Preschool Inventory) + .05 (TOBE-Math) + .03 (DAD) - 37.88.
MAT - Word Discrimination	.66	.63 (MRT)* + .29 (TOBE-Language)* + .15 (DAC)* + .15 (Lee-Clark)* + .05 (DAD) + .05 (TOBE-Math) + .002 (Preschool Inventory) - 33.67.
MAT - Reading	.71	.70 (MRT)* + .49 (TOBE-Language)* + .23 (DAC)* + .16 (Lee-Clark)* + .01 (Preschool Inventory) + .01 (DAD) + .002 (TOBE-Math) - 40.78.
MAT - Arithmetic	.73	.48 (Preschool Inventory)* + .32 (MRT)* + .21 (TOBE-Language)* + .12 (Lee-Clark)* + .07 (TOBE-Math) + .003 (DAC) + .001 (DAD) - 46.14.

*variables contributing at a statiscially signifianct level ($p \leq .05$) to the prediction of the dependent variable

Table 4

Analysis of Variance Summary Tables for Analysis of Residuals Based on TOBE-Language and Lee-Clark as Predictors

I. MAT - Word Discrimination		R = .58	
Source	df	MS	F
Race	1	105.11	1.63
Sex	1	427.39	6.64*
Race x Sex	1	138.20	2.15
Error	318	64.40	
II. MAT - Word Knowledge		R = .58	
Source	df	MS	F
Race	1	19.01	0.30
Sex	1	746.70	11.90*
Race x Sex	1	220.33	3.51
Error	318	62.71	
III. MAT - Reading		R = .63	
Source	df	MS	F
Race	1	79.68	1.35
Sex	1	244.80	4.14
Race x Sex	1	80.22	1.36
Error	318	59.09	
IV. MAT - Arithmetic		R = .68	
Source	df	MS	F
Race	1	1.51	0.03
Sex	1	75.14	1.42
Race x Sex	1	130.48	2.47
Error	318	52.91	

* significant at $p \leq .01$

References

ANASTASI, A. Psychological Testing. 4th edition. New York: MacMillan, 1976.

BERNAL, E. M. A response to "Educational uses of tests with disadvantaged students." American Psychologist, 1975, 30, 93-95.

CARDALL, C., & COFFMAN, [illegible]. A method for comparing the performance of different groups on the items in a test. Research and Development Reports, 1964, 64-5 No. 9, College Entrance Examination Board.

CLEARY, T. A. Test bias: Prediction of grades of Negro and White students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

CLEARY, T. A., HUMPHREYS, L. G., KENDRICK, S. A., & WESMAN, A. Educational uses of tests with disadvantaged students. American Psychologist, 1975, 30, 15-41.

COCHRAN, W. G., & COX, G. M. Experimental designs. Wiley, 1957.

EDUCATIONAL TESTING SERVICE, Preschool Inventory-Revised Edition. Princeton, 1970.

HARRINGTON, G. M. Intelligence tests may favour the majority groups in a population. Nature, 1975, 258, 708-709.

HILDRETH, G. H., GRIFFITHS, N. C., & MCGAUVRAN, M. E. Metropolitan Readiness Tests. New York: Harcourt, Brace, Jovanovich, Inc., 1969.

KATZENMEYER, W. G., & STENNER, A. J. Estimation of the invariance of factor structures across sex and race with implications for hypothesis testing. Educational and Psychological Measurement, 1977, 37, 111-119.

LEE, J. M., & CLARK, W. L. Lee-Clark Reading Readiness Test. Monterey, California: California Test Bureau, 1962.

LINN, R. L., & WERTS, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

MCCARTHY, D. McCarthy Scales of Children's Abilities. New York: Psychological Corporation, 1972.

MITCHELL, B. C. Predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for white and for negro pupils. Educational and Psychological Measurement, 1967, 27, 1047-1054.

MOSS, M. H. Tests of Basic Experiences. Monterey, California CTB/McGraw-Hill, 1970.

OAKLAND, T., & MATUSZEK, P. Using tests in non-discriminatory assessment. In T. Oakland (Ed.) Psychological and educational assessment of minority group children. New York: Brunner/Mazel, 1977.

OZEHOSKY, R. J., & CLARK, E. T. Verbal and non-verbal measures of self-concept among kindergarten boys and girls. Psychological Reports, 1971, 28, 195-199.

POTTHOFF, R. F. Statistical aspects of the problem of biases in psychological tests. Institute of Statistics Mimeo Series No. 479, Chapel Hill, NC: Department of Statistics, 1966.

REYNOLDS, C. R. Differential validity of several preschool assessment instruments for blacks, whites, males, and females. Unpublished doctoral dissertation, Athens Georgia: University of Georgia, 1978. a.

REYNOLDS, C. R. The McCarthy Drawing Tests as a group instrument. Contemporary Educational Psychology, 1978, 3, 169-174. b.

REYNOLDS, C. R. A factor analytic study of the Metropolitan Readiness Tests. Contemporary Educational Psychology, 1979, 4, 315-317.

REYNOLDS, C. R. The invariance of the factorial validity of the Metropolitan Readiness Tests for blacks, whites, males, and females. Educational and Psychological Measurement, 1979, in press.

REYNOLDS, C. R. Should we screen preschoolers? Contemporary Educational Psychology, 1979, 4, 175-181.

REYNOLDS, C. R., & HARTLAGE, L. C. Comparison of WISC and WISC-R racial regression lines for academic prediction with black and white referred children. Journal of Consulting and Clinical Psychology, 1979, 47, 589-591.

THORNDIKE, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Notes

[1]Larry P. et al v. Riles et al., 343 F. Supp. 1306 (D.C.N.D. Cal., June 20, 1972).

[2]This procedure is conceptually similar to analysis of covariance and tends to provide slightly inflated F-ratios (see Cochran & Cox, 1957, for a statistical explanation) thus giving a "conservative" test of the question of bias. That is, this procedure is more likely to overestimate the presence of bias than are more exact, but computationally complex, procedures.